P-ISSN: 2548-9704
E-ISSN: 2686-0880

PENERAPAN TEKNIK SAMPLING UNTUK MENGATASI *IMBALANCE CLASS* PADA KLASIFIKASI *ONLINE SHOPPERS INTENTION*

Ardiyansyah[1)], Panny Agustia Rahayuningsih[,2)]

[1, 2] Program Studi Sistem Informasi Akuntansi Kampus Kota Pontianak, Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika
Jl. Abdurrahman Saleh No. 18 A, Pontianak, Indonesia
E-mail : Ardiyansyah.arq@bsi.ac.id, panny.par@bsi.ac.id

ABSTRACT

Online shopping or e-commerce is a transaction process carried out through intermediary media in the form of online trading sites and social media that provide goods and services that are traded. Much research has focused on predicting realtime income for shopping web sites. The dataset consists of 10 numeric attributes and 8 category attributes. In this dataset, there is a possibility that there is an unbalanced target variable. Where, this is the case for each individual target value in the dataset. online shopper inttention aims to predict whether users generate revenue or not. Class imbalance occurs when the minority class is smaller than the majority class. Using unbalanced data will result in a minority class producing low accuracy values. Sampling methods are SMOTE, Undersampling and Oversampling To overcome the problem of class imbalance (imbalance class) as a measurement of performance. whereas, the classification algorithm method used is random forest, KNN, and Naive Bayes. The results of the evaluation and validation, it can be concluded that the best sampling method in overcoming the imbalance class in this study is the oversampling method. The random forest model without sampling has the highest f-measure value than the other models, which is 0.898. After applying the sampling method, the results of the comparison between the smote + random forest, undersampling + random forest and oversampling + random forest models. The best model with the highest f-measure and AUC is the oversampling + random forest model, the f-measure is 0.976 or 98% and the AUC value is 0.998. So the oversampling + random forest model is the best model in the study of the application of sampling techniques in overcoming the imbalance class in the online shopper intention enthusiast classification.

Keywords: ***online shopper intention,e-commerce, imbalance class, classification***

ABSTRAK

Belanja online atau *e-commerce* adalah sebuah proses transaksi yang dilakukan melalui media perantara yaitu berupa situs-situ jual beli online dan media sosial yang menyediakan barang dan jasa yang diperjualbelikan. Banyak penelitian memfokuskan untuk memprediksi pendapatan *realtime* untuk situs web belanja. Dataset terdiri dari 10 atribut numerik dan 8 atribut kategori.Pada dataset ini, terdapat kemungkinan ada variabel target yang tidak seimbang. Dimana, hal ini menjadi kasus pada masing-masing nilai target individu dalam dataset. *online shopper inttention* bertujuan untuk memprediksi apakah pengguna menghasilkan pendapatan atau tidak. Ketidak seimbangan kelas terjadi ketika kelas minoritas lebih kecil dari kelas mayoritas. Dengan menggunakan data yang tidak seimbang akan mengakibatkan kelas minoritas menghasikan nilai akurasi yang rendah.metode sampling yaitu SMOTE, *Undersampling* dan *Oversampling* Untuk mengatasi masalah ketidakseimbangan kelas (*imbalance class*) sebagai pengukuran kinerjanya. sedangkan, Metode algoritma pengklasifikasian yang digunakan adalah *random forest*, KNN, dan Naive Bayes. Hasil dari evaluasi dan validasi, dapat disimpulkan

P-ISSN: 2548-9704
E-ISSN: 2686-0880

bahwa metode sampling yang terbaik dalam mengatasi *imbalance class* pada penelitian ini adalah metode *oversampling*. Model *random forest* tanpa sampling memiliki nilai *f-measure* tertinggi dari model lainnya yaitu sebesar 0.898. Setelah menerapkan metode sampling, hasil dari perbandingan antara model smote+random forest, *undersampling+random forest* dan *oversampling+random forest*. Model terbaik dengan memiliki nilai f-measure dan AUC tertinggi adalah model *oversampling+random forest* yaitu nilai f-*measure* sebesar 0.976 atau 98% dan nilai AUC sebesar 0.998. Sehingga model *oversampling+random forest* merupakan model terbaik dalam penelitian penerapan Teknik sampling dalam mengatasi *imbalance class* pada klasifikasi peminat *online shopper intention.*

Kata kunci: *online shopper intention. e-commerce, imbalance class*, klasifikasi.

I. PENDAHULUAN

Dalam beberapa tahun terakhir ini, internet telah berkembangan sangat pesat. Internet saat ini tidak hanya menjadi media informasi atau komunikasi saja. Internet telah muncul sebagai alat media pemasaran yang berguna untuk menambahkan pendapatan seseorang dalam perekonomian dan dapat memenuhi kebutuhan keinginan masyarakat dalam berbelanjaan *online* dengan lebih praktis. Banyak perusahan yang memanfaatkan internet sebagai media pemasaran produknya. Belanja online adalah fenomena yang berkembang sangat pesat saat ini [1]. Belanja online atau *e-commerce* adalah sebuah proses transaksi yang dilakukan melalui media perantara yaitu berupa situs-situ jual beli online dan media sosial yang menyediakan barang dan jasa yang diperjualbelikan [2]. Pada situs web medium.com analytic vidhya menjabarkan bahwa pengguna yang masuk ke dalam situs web belanja online,mengetahui apakah pengguna tersebut akan melakukan pembelian atau tidak memiliki nilai ekonomi yang besar. banyak penelitian memfokuskan untuk memprediksi pendapatan realtime untuk situs web belanja. Tujuan dari belanja online adalah untuk menilai dan menganalisa perilaku niat belanja konsumen. Sebelum melakukan pembelian, konsumen akan mengumpulkan informasi tentang produk yang akan menjadi pertimbangan [3] baik itu menggunakan situs website online maupun aplikasi toko online yang akan memberikan keuntungan pada toko tersebut dan bagiamana toko online dapat mempertahankan pelanggannya agar tetap belanja di toko online sehingga lebih diminati dan disukai pembelinya [2].

Online shoppers intention merupakan *dataset* yang dapat digunakan untuk membangun model pembelajaran mesin prediktif yang dapat mengkategorikan pengguna sebagai, menghasilkan pendapatan dan non-pendapatan berdasarkan perilaku mereka saat menavigasi situs web. *Dataset* ini diperoleh dari UCI *Machine Learning repository*. *Dataset* terdiri dari vektor fitur yang dimiliki hingga 12.330 sesi. *Dataset* dibentuk sehingga setiap sesi akan menjadi milik pengguna yang berbeda dalam periode 1 tahun untuk menghindari kecenderungan kampanye tertentu, hari khusus, profil pengguna, atau periode. Dataset terdiri dari 10 atribut numerik dan 8 atribut kategori. Atribut pendapatan dapat digunakan sebagai label kelas. File ini terdiri dari berbagai Informasi yang berkaitan dengan perilaku pelanggan di situs web belanja online. Pada *dataset* ini, salah satu yang perlu diperhatikan adalah bahwa ada kejadian dalam variabel target dataset mungkin tidak seimbang. Nilai target *dataset* ini menunjukkan apakah *dataset* seimbang atau tidak seimbang. Dimana, hal ini menjadi kasus pada masing-masing nilai target *individu* dalam *dataset*. *online shopper inttention* bertujuan untuk memprediksi apakah pengguna menghasilkan pendapatan atau tidak.
Salah satu cara untuk mendapatkan informasi atau pola dari kumpulan data yang besar adalah dengan menggunakan teknik-teknik dalam data mining (Ardiyansyah et al., 2018). Data mining merupakan sebuah proses

P-ISSN: 2548-9704
E-ISSN: 2686-0880

ekstraksi untuk mendapatkan suatu informasi yang sebelumnya tidak diketahui dari sebuah data [4]. Jikat dilihat dari *online shoppers intentio*n yang digunakan, secara umum *dataset* bersifat tidak seimbang (*imbalanced*) . klasifikasi data pembagian yang tidak seimbang dapat menurunkan kinerja algoritma belajar (*learning algorithm*) pengklasifikasian standar secara signifikan [5]. Klasifikasi merupakan hal yang sangat penting untuk melakukan secara akurat memprediksi perilaku penggunaan konsumen atau niat belanja *online* pembelinya. Algoritma Klasifikasi banyak digunakan untuk menentukan data mining algoritma mana yang dianggap paling baik dalam melakukan proses pengklasifikasian [6]. Ketidak seimbangan kelas terjadi ketika kelas minoritas lebih kecil dari kelas mayoritas. Dengan menggunakan data yang tidak seimbang akan mengakibatkan kelas minoritas menghasikan nilai akurasi yang rendah. Ada tiga pendekatan untuk menangani dataset yang tidak seimbang: level data, level algoritmik, dan metode penggabungan atau ensemble metode. Pendekatan teknik level data yang berusaha menyeimbangkan distribusi data dengan metode *over-sampling* dan *undersampling*. Pendekatan level algoritmik yaitu dengan mengembangkan algoritma baru atau memodifikasi metode yang ada untuk memperhitungkan arti dari kelas minor. Sedangkan Ketiga dengan mengkombinasikan pendekatan algoritma dan pendekatan level data [7]. Untuk mengatasi masalah ketidakseimbangan kelas (*imbalance class*), penelitian ini melakukan pengukuran kinerja untuk mengatasi masalah *imbalance class* menggunakan pendekatan level data dengan menerapkan metode sampling yaitu SMOTE, *Undersampling* dan *Oversampling* serta Metode algoritma pengklasifikasian yang digunakan adalah *random forest*, KNN, dan Naive Bayes sehingga dapat menghasilkan nilai akurasi dan pendekatan level data terbaik dalam menyelesaikan ketidakseimbangan kelas pada *online shopper intention*.

2. METODOLOGI PENELITIAN

Pada penelitian ini, penulis menggunakan weka 3.8.3 sebagai *tools* yang digunakan untuk pengujian model. sedangkan dataset yang digunakan adalah data *online shoppers intention* dari *uci machine learning repository* tahun 2019. metode sampling yang digunakan dalam mengatasi masalah *imbalance class* adalah smote, undersampling dan oversampling. selain itu, ada tiga metode klasifikasi yang digunakan pada penelitian ini yaitu knn, naïve bayes dan random forest.

Metode yang digunakan pada penelitian ini adalah eksperimen. Penelitian eksperimen adalah penelitian yang melibatkan investigasi sebab akibat dengan menggunakan uji coba yang dikontrol sendiri. Cukup sering penelitian semieksperimental mendapatkan kendala karena kurangnya akses terhadap sampel, masalah etika dan sebagainya. Eksperimen biasanya dilakukan pada pengembangan, evaluasi dan pemecahan masalah proyek. [8]

Berikut adalah tahapan penelitian yang dilakukan, diantaranya [8]:

A. *Data Gathering*
   Pada tahap pertama dijelaskan mengenai dataset yang digunakan pada penelitian ini, dataset yang digunakan adalah data *online shoppers intention* dari *UCI Machine Learning Repository* tahun 2019.
B. *Data Pre-processing*
   Tahap yang kedua, menerapkan metode resampling untuk mengatasi *imbalance class* pada dataset, metode yang digunakan seperti SMOTE, Undersampling dan Oversampling.
C. *Proposed Method*
   Metode yang digunakan berdasarkan masalah dan sesuai dengan metode klasifikasi yang pernah digunakan oleh para peneliti yaitu KNN, Naïve Bayes dan Random Forest.
D. *Result Evaluation and Validation*
   Pada tahap terakhir, akan di lakukan pengujian model dengan membandingkan antara tiga metode klasifikasi yang dilakukan tanpa sampling, dengan SMOTE, Undersampling dan Oversampling, dengan menggunakan data *training* dan data *testing*

P-ISSN: 2548-9704
E-ISSN: 2686-0880

yang divalidasi dengan *k-fold cross validation* yaitu 10-*fold cross validation*. Diman 10-*fold cross validation* data akan dibagi menjadi 10 bagian yaitu 9 data training dan 1 data testing. Sedangkan untuk mengukur kinerja masing-masing algoritma dilakukan dengan menggunakan *confusion matrix* dan di ukur berdasarkan nilai *F-Measure* dan AUC.

Berikut adalah tabel *Confusion Matrix* [9] yang digunakan pada penelitian ini:

Tabel 1. Tabel *Confusion Matrix*

	Predicted negative	*Predicted Positive*
Actual Negative	TN	FP
Actual Positive	FN	TP

Confusion matrix adalah alat untuk menganalisa seberapa baik kinerja dari pengklasifikasi dapat mengenali tupel dari kelas yang berbeda [10]. Confusion matrix memberikan penilaian kinerja klasifikasi yang digambarkan dengan perbandingan antara hasil prediksi dengan kenyataan [10].
Dari tabel model *confusion matrix* diatas didapat persamaan sebagai berikut :

Precision = $\frac{TP}{TP+FP}$ [11]

Recall = $\frac{TP}{TP+FN}$ [11]

F-Measure = $2\frac{Precision.\ Recall}{Precision + Recall}$ [11]

Nilai recall dan precision menjadi tolak ukur dalam penentuan nilai f-measure, sehingga dapat menghasilkan metrik yang efektif dalam mengatasi masalah ketidakseimbangan kelas.

Yang terjadi pada masalah imbalanced class adalah kelas *minority* lebih mendominasi dibandingkan kelas *majority*. Sehingga penggunaan Area Under ROC Curve (AUC) dapat digunakan untuk memberikan metrik numerik single untuk dapat membandingkan kinerja dari model, nilai AUC berkisar dari 0 sampai 1 dan model yang lebih baik prediksinya adalah yang mendekati nilai 1 [10].

AUC = $\frac{1+TP-FP}{2}$ [10]

Kurva ROC adalah grafik antara sensitivitas (*true positive rate*) pada sumbu Y dengan 1-spesifisitas pada sumbu X (*false positive rate*), curve ROC ini seakan-akan menggambarkan tawar-menawar antara sumbu Y atau sensitivitas dengan sumbu X atau spesifisitas. Nilai dalam kurva ROC dapat menjadi evaluasi sehingga dapat membandingkan algoritma. Kurva ROC merupakan teknik untuk memvisualisasikan, mengatur dan memilih klasifikasi berdasarkan kinerja dari algoritma [10].

3. HASIL DAN PEMBAHASAN

Pada bagian ini, pembahasan hasil pengujian dilakukan dengan menggunakan WEKA 3.8.3. Kemudian dataset yang digunakan yaitu online shoppers intention dengan 18 atribut, 12000 instance dan 2 class yaitu class “True” dan class “False”. Pengujian pertama dilakukan dengan menggunakan metode KNN, Naïve Bayes dan Random Forest tanpa Pre-Processing atau tanpa Teknik sampling. Kemudian ketiga metode klasifikasi tersebut diuji kembali dengan data yang di Pre-Processing terlebih dahulu dengan menggunakan Teknik sampling yaitu dengan metode SMOTE, Under-Sampling dan Over-Sampling. Terakhir, hasil dari pengujian tersebut akan dibandingkan untuk menemukan Teknik sampling yang lebih baik dalam mengatasi imbalance class dan untuk mengklasifikasi peminat online shoppers.

3.1. Hasil Pengujian Tanpa Sampling

Pada tahap pertama, pengujian dilakukan tanpa menggunakan metode sampling dan langsung dilakukan pengujian dengan menggunakan metode KNN, Naïve Bayes dan Random Forest. Hasil pengujian tanpa sampling dapat dilihat di tabel 2.

P-ISSN: 2548-9704
E-ISSN: 2686-0880

Tabel 2. Hasil Perbandingan Metode Klasifikasi Tanpa Sampling

Algoritma	*TP-Rate*	*FP-Rate*	*Precission*	*Recall*	*F-Measure*
KNN	**0.854**	**0.714**	**0.825**	**0.854**	**0.816**
Naïve Bayes	**0.816**	**0.306**	**0.856**	**0.816**	**0.830**
Random Forest	**0.902**	**0.356**	**0.896**	**0.902**	**0.898**

Pada tabel 2 menunjukan bahwa perbandingan antara 3 metode klasifikasi yang dilakukan tanpa sampling tersebut, yang memiliki nilai f-Measure yang lebih baik adalah metode Random Forest dengan nilai f-measure 0.898 dibandingkan dengan metode Naïve Bayes yang hanya memiliki nilai f-measure sebesar 0.830 dan KNN sebesar 0.816.

3.2. Hasil Pengujian dengan SMOTE

Pada bagian ini, pengujian dilakukan pada 3 metode klasifikasi yaitu KNN, Naïve Bayes dan Ranom Forest dengan menerapkan metode SMOTE (Synthetic Minority Over-sampling Technique). Hasil pengujian dengan menerapkan metode SMOTE dapat dilihat pada tabel 3.

Tabel 3. Hasil Perbandingan Metode Klasifikasi Dengan Metode SMOTE

Algoritma	*TP-Rate*	*FP-Rate*	*Precission*	*Recall*	*F-Measure*
SMOTE +KNN	**0.786**	**0.352**	**0.782**	**0.786**	**0.784**
SMOTE +Naïve Bayes	**0.765**	**0.203**	**0.816**	**0.765**	**0.777**
SMOTE +Random Forest	**0.908**	**0.149**	**0.907**	**0.908**	**0.907**

Pada tabel 3 menunjukan bahwa setelah menerapkan metode SMOTE, terjadi peningkatan nilai f-measure pada metode random forest dengan peningkatan sebesar 1%. dan penurunan terjadi pada metode KNN yaitu f-measure sebesar 3%. dan Naïve Bayes 5%. Dari hasil tersebut, dapat diketahui bahwa nilai f-measure dari model SMOTE + Random Forest sebesar 0.907 lebih tinggi dibandingkan dengan model SMOTE+KNN sebesar 0.784 dan SMOTE+Navie Bayes yaitu sebesar 0.777.

Sehingga apabila hasil tersebut dibandingkan maka metode SMOTE+random forest lebih baik dari metode SMOTE+knn, SMOTE+naïve bayes dan metode random forest tanpa sampling.

3.3. Hasil Pengujian dengan Undersampling

Pada tahap ini, pengujian dilakukan dengan menerapkan metode undersampling. Hasil pengujian dengan menerapkan metode undersampling dapat dilihat pada tabel 4.

Tabel 4. Hasil Perbandingan Metode Klasifikasi Dengan Metode Undersampling

Algoritma	*TP-Rate*	*FP-Rate*	*Precission*	*Recall*	*F-Measure*
Undersampling + KNN	**0.892**	**0.108**	**0.896**	**0.892**	**0.892**
Undersampling + Naïve Bayes	**0.760**	**0.240**	**0.768**	**0.760**	**0.758**
Undersampling + Random Forest	**0.947**	**0.053**	**0.948**	**0.947**	**0.947**

Pada tabel 4 menunjukan bahwa setelah menerapkan metode undersampling, terjadi peningkatan nilai f-measure pada metode random forest, yang sebelumnya pada saat menerapkan metode smote hanya meningkat 1%, dan pada saat menerapkan metode undersampling meningkat menjadi 4%. Sedangkan metode KNN, pada saat menerapkan metode smote mengalami penurunan sebesar 3%, tetapi ketika menerapkan metode undersampling mengalami peningkatan sebesar 8%. Berbeda dengan metode Naïve Bayes yang mengalami penurunan baik pada penerapan metode smote sebesar 5% dan pada penerapan metode undersampling sebesar 7%.

Dari hasil tersebut, dapat diketahui bahwa nilai f-measure dari model Undersampling+Random Forest sebesar 0.947 lebih tinggi dibandingkan dengan model

P-ISSN: 2548-9704
E-ISSN: 2686-0880

Undersampling+KNN sebesar 0.892 dan Undersampling+Navie Bayes yaitu sebesar 0.758.

Sehingga apabila hasil tersebut dibandingkan maka metode Undersampling+Random Forest lebih baik dari metode Undersampling+KNN dan Undersampling+Navie Bayes. Bahkan jika dibandingkan dengan metode SMOTE+random forest, metode undersampling+random forest sebesar 0.947 lebih unggul dari metode SMOTE+Random Forest sebesar 0.907, dan mengalami peningkatan sebesar 4%.

3.4. Hasil Pengujian dengan Oversampling

Pada tahap ini, pengujian dilakukan dengan menerapkan metode oversampling. Hasil pengujian dengan menerapkan metode oversampling dapat dilihat pada tabel 5.

Tabel 5. Hasil Perbandingan Metode Klasifikasi Dengan Metode Oversampling

Algoritma	*TP-Rate*	*FP-Rate*	*Precission*	*Recall*	*F-Measure*
Oversampling+ KNN	**0.954**	**0.046**	**0.956**	**0.954**	**0.954**
Oversampling+ Naïve Bayes	**0.763**	**0.237**	**0.775**	**0.763**	**0.760**
Oversampling+ Random Forest	**0.976**	**0.024**	**0.977**	**0.976**	**0.976**

Pada tabel 5 menunjukan bahwa setelah menerapkan metode oversampling, terjadi peningkatan nilai f-measure pada metode random forest, yang sebelumnya pada saat menerapkan metode smote hanya meningkat 1%, metode undersampling meningkat 4% dan pada saat menerapkan metode oversampling meningkat menjadi 8%. Sedangkan metode KNN, pada saat menerapkan metode smote mengalami penurunan sebesar 3%, menerapkan metode undersampling meningkat sebesar 8% dan pada saat menerapkan metode oversampling meningkat menjadi 14%. Berbeda dengan metode Naïve Bayes yang mengalami penurunan baik pada penerapan metode smote sebesar 5% dan pada penerapan metode undersampling dan oversampling sebesar 7%.

Dari hasil tersebut, dapat diketahui bahwa nilai f-measure dari model Oversampling+Random Forest sebesar 0.976 lebih tinggi dibandingkan dengan model Oversampling+KNN sebesar 0.954 dan Oversampling+Navie Bayes yaitu sebesar 0.760.

Sehingga apabila hasil tersebut dibandingkan maka metode Oversampling+Random Forest lebih baik dari metode Oversampling+KNN dan Oversampling+Navie Bayes. Bahkan jika dibandingkan dengan metode SMOTE+random forest dan metode undersampling+random forest, maka metode Oversampling lebih unggul dari kedua metode tersebut, dan lebih meningkat 4% dari peningkatan sebelumnya.

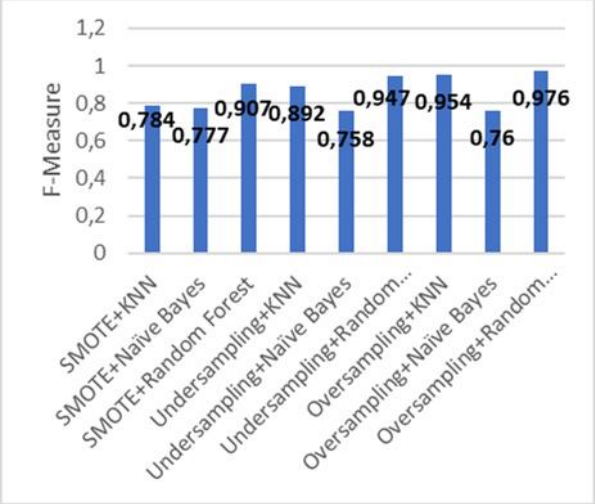
3.5. Hasil Perbandingan Penerapan Metode Sampling pada Metode klasifikasi

Pada bagian ini, akan ditampilkan hasil dari pengujian metode sampling yaitu metode smote, undersampling dan oversampling. Kemudian akan dibandingkan untuk mendapatkan model terbaik dari perbandingan dengan melihat dari pengukuran nilai f-measure dan AUC.

Berikut grafik perbandingan penerapan metode sampling pada metode klasifikasi

P-ISSN: 2548-9704
E-ISSN: 2686-0880

dengan pengukuran f-measure yang dapat dilihat pada gambar 1 dibawah ini.



Gambar 1. Hasil Perbandingan Dengan Pengukuran F-*Measure*

Pada gambar 1 menunjukan bahwa Pada hasil pengujian dengan pengukuran f-measure yang telah dilakukan dan disajikan pada grafik diatas, menunjukan bahwa metode sampling yang terbaik pada penelitian adalah metode Oversampling, mengungguli metode Undersampling dan SMOTE. Hal ini ditunjukan bahwa adanya peningkatan nilai f-measure pada ketiga model klasifikasi yang diuji. Selain itu, model klasifikasi yang terbaik pada penelitian ini adalah model oversampling+random forest yang memiliki nilai f-measure yang lebih baik dari model yang lainnya yaitu dengan nilai f-measure sebesar 0.976 atau 98% dengan peningkatan sebesar 0.029 atau 3%.

Pada tabel 6 menunjukan bahwa pada hasil pengujian dengan pengukuran AUC yang telah dilakukan menunjukan bahwa penerapan metode sampling pada metode klasifikasi yang terbaik pada penelitian ini adalah model oversampling+random forest dengan nilai AUC sebesar 0.998. model terbaik kedua adalah model undersampling+randm forest, diikuti dengan model smote+random forest, oversampling+knn, undersampling+knn, smote+naïve bayes, oversampling+naïve bayes, undersampling+naïve bayes dan model dengan performa kurang baik adalah smote+knn.

Berikut hasil perbandingan penerapan metode sampling pada metode klasifikasi dengan pengukuran AUC yang sajikan pada tabel 6 dibawah ini.

Tabel 6. Hasil Perbandingan *Metode Sampling* dengan AUC

Algoritma	***AUC***
SMOTE+KNN	0.720
SMOTE+Naïve Bayes	0.845
SMOTE+Random Forest	0.960
Undersampling+KNN	0.892
Undersampling+Naïve Bayes	0.834
Undersampling+Random Forest	0.990
Oversampling+KNN	0.955
Oversampling+Naïve Bayes	0.840
Oversampling+Random Forest	0.998

4. KESIMPULAN

Berdasarkan hasil pengujian dengan menerapkan metode sampling pada metode klasifikasi yang dilakukan pada dataset *online shoppers intention.* Berdasarkan hasil dari evaluasi dan validasi, dapat disimpulkan bahwa metode sampling yang terbaik dalam mengatasi *imbalance class* pada penelitian ini adalah metode oversampling. Dan metode klasifikasi terbaik dengan penerapan metode sampling adalah metode random forest. Hasil klasifikasi menunjkan bahwa model random forest tanpa sampling memiliki nilai f-measure tertinggi dari model lainnya yaitu sebesar 0.898. Setelah menerapkan metode sampling, hasil dari perbandingan antara model smote+random forest, undersampling+random forest dan oversampling+random forest. Model terbaik dengan memiliki nilai f-measure dan AUC tertinggi adalah model oversampling+random forest yaitu nilai f-measure sebesar 0.976 atau 98% dan nilai AUC sebesar 0.998. Sehingga model oversampling+random forest merupakan model terbaik dalam penelitian penerapan

P-ISSN: 2548-9704
E-ISSN: 2686-0880

Teknik sampling dalam mengatasi imbalance class pada klasifikasi peminat *online shoppers*.

5. SARAN

Pada penelitian selanjutnya, penelitian ini dapat dikembangkan untuk menghasilkan model yang lebih baik, seperti berikut.

a. Pada penelitian selanjutnya dapat menggunakan metode yang berbeda dalam mengatasi masalah *imbalance class*.
b. Pada penelitian selanjutnya dapat menggunakan dataset yang memiliki ketidakseimbangan kelas dengan tingkat lebih tinggi.
c. Dapat menggunakan metode klasifikasi yang berbeda.

DAFTAR PUSTAKA

[1] Y. J. Lim, A. Osman, S. N. Salahuddin, A. R. Romle, and S. Abdullah, "Factors Influencing Online Shopping Behavior: The Mediating Role of Purchase Intention," *Procedia Econ. Financ.*, vol. 35, no. October 2015, pp. 401–410, 2016.

[2] D. A. Harahap, "Perilaku Belanja Online Di Indonesia: Studi Kasus," *JRMSI - J. Ris. Manaj. Sains Indones.*, vol. 9, no. 2, pp. 193–213, 2018.

[3] A. A. Mahardika and Saino, "Ayu Anastasya Mahardika dan Saino; Analis Faktor Yang Mempengarruhi … 917," *J. Ilmu Manaj.* |, vol. 2, no. 1996, 2014.

[4] Ardiyansyah, P. A. Rahayuningsih, and Reza Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," *J. Khatulistiwa Inform.*, vol. VI, no. 6, pp. 20–28, 2018.

[5] A. Saifudin and S. Wahono, "Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 2, pp. 76–85, 2015.

[6] P. A. Rahayuningsih, "Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Tingkat Kematian Dini Kanker dengan Dataset Early Death Cancer," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 4, no. 2, p. 63, 2019.

[7] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," *J. Inform.*, vol. 5, no. 2, pp. 175–185, 2018.

[8] C. W. Dawson, *Projects in Computing and Information Systems A Student's Guide*, 2nd ed., vol. 2. England: Pearson Education, 2009.

[9] H. He and Y. Ma, *Imbalanced Learning - Foundations, Algorithms, and Applications*, 1st ed. New Jersey: The Institute of Electrical and Electronics Engineers, Inc, 2013.

[10] Fitriyani and R. S. Wahono, "Integrasi Bagging dan Greedy Forward Selection pada Prediksi Cacat Software dengan Menggunakan Naïve Bayes," *J. Softw. Eng.*, vol. 1, no. 2, pp. 101–108, 2015.

[11] Lior Rokach and Oded Maimon, *DATA MINING WITH DECISION TREES Theory and Applications*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2015.

P-ISSN: 2548-9704
E-ISSN: 2686-0880